THE PREDICTIVE SAMPLE REUSE METHOD

by

Seymour Geisser

Technical Report No. 226

December, 1973

University of Minnesota

Minneapolis, Minnesota

1. Introduction.

A highly flexible synthesis of two time honored and widely used methods in data analysis namely cross-validation and fitting approaches are presented with a crucial change in emphasis. The thrust here is essentially that the prediction of observables or potential observables is of much greater relevance than the estimation of what are often artificial constructs - parameters. This new approach is designated as the method of Predictive Sample Reuse, Geisser (1974).

Basically this method was also independently and simultaneously propounded by M. Stone (1974) who termed it the Cross-validatory method for what may be historical reasons. There are some other slight differences in terminology which are relatively unimportant e.g., Stone uses a loss function and here it is termed a discrepancy function in order to segregate it from Decision Theory notions and keep it within the realm of Data Analysis. Stone's presentation involves single observational omissions and also proceeds in two other directions which he terms "double cross" and "model mix" allowing for greater generality. My own development allows for multiple observational omissions and also yields a desirable degree of flexibility as well as a close congruence with the purposes of particular prediction. Stone also provides an exhaustive historical background for the approach stressing in particular the importance of a seminal paper by Mosteller and Tukey [1968], a paper,

unfortunately, that I was unaware of when I devised this approach. For me this approach arose from the strong feeling that prediction was more relevant for inference than parameter estimation, see Geisser [1971], and that prediction, unlike parameter estimation, could often be adequately assessed in real situations. Also it was felt that the Bayesian approach which certainly would lead to the desirable goal of predictive distributions, was too highly structured, too rich and too parameter bound for many situations that arise in practice. In any case easily understandable and workable data featuring tools are attractive additional alternatives to a statistician's armamentarium.

2. Predictive Sample Reuse Methodology.

Suppose we have a sample of size N of data $X = (x_1,\ldots,x_N)$ each x_i associated with a known y_i, $Y = (y_1,\ldots,y_N)$. Further we are interested in predicting a new value x for a corresponding known value of y without necessarily involving distributional or parametric assumptions. Suppose a predictive function is chosen and specified symbolically as

$$x = x(X,Y.y;\mu) \quad \mu \in \Omega \tag{2.1}$$

where μ is some set of unknown values. It must be stressed that in this approach μ is not a platonic ideal nor in any sense a true value of paramount importance. It is to be regarded as merely a convenient way of forming a predictive function. Let $X_i^{(N-n)}$ represent the i^{th} partition of the sample into $N-n$ retained and n omitted observations $0 < n \leq M$, where M is the largest integer such that the predictive function (2.1) can be formed with $N-M$ observations. The observational set X is partitioned such that $X_i^{(N-n)} = (x_{ir}^{(N-n)}, x_{i0}^{(n)})$ with corresponding partition $Y_i^{(N-n)} = (y_{ir}^{(N-n)}, y_{i0}^{(n)})$ ($x_{ir}^{(N-n)}$ and $x_{i0}^{(n)}$ are the sets of $N-n$ retained and n omitted observations respectively) is the i^{th} partition belonging to a set Γ of partitions relevant to a particular schema of observational omissions. Let the total number of such partitions be $P(N,n,\Gamma)$, or simply P. The specified predictive function is then applied to the retained observations for prediction of the omitted observations for each partition with the unknown value μ estimated by means of optimizing an average discrepancy function, say

$$D_{N,n}(\mu) = P^{-1}{}_{n}{}^{-1} \sum_{i \in \Gamma} d(x_{i0}^{(n)}, \hat{x}_{i0}^{(n)}(x_{ir}^{(N-n)}, Y_i^{(n-n)}; \mu)) \qquad (2.2)$$

where each observation in the set $\hat{x}_{i0}^{(n)}$ is of the form of the predictive function and d is a measure of the discrepancy of $x_{i0}^{(n)}$ from $\hat{x}_{i0}$. $D_{N,n}(\mu)$ is then optimized with respect to μ in some sense. On the basis that this leads to a solution say, $\hat{\mu}$, we obtain the predictor $\hat{x} = x(X,Y,y;\hat{\mu})$.

Aside from the choice of a discrepancy measure, there arise several other questions, namely:

(1) How do we assess various different predictive functions for a given schema of partitions?

(2) How do we choose for a given predictive function and data set alternative schemata of partitioning?

(3) Lastly, once we have decided jointly on a predictive function and a schema for partitions, how do we attach a measure of discrepancy to the predictor?

Only the first question appears to have a semblance of a clear cut answer. For a given discrepancy measure we could consider for the i^{th} partition the set of retained observations $x_{ir}^{(N-n)}$ and partition this into two sets $x_{ir}^{(N-2n)} = (x_{irr}^{(N-2n)}, x_{ir0}^{(n)})$. On this reduced set of $N-n$ observations we would, in the same manner as previously, obtain a $\hat{\mu}_i$. Repeating this for each i we would then compute an overall discrepancy measure (not necessarily based on the same d as was used to attain the predictor)

$$D^{*}_{N-n} = P^{-1}{}_{n}{}^{-1} \sum_{i\epsilon\Gamma} d(x^{(n)}_{iO}, \hat{x}^{(n)}_{iO}(x^{(N-n)}_{ir}, Y^{(N-n)}_{i}; \hat{\mu}_{i}) \qquad (2.3)$$

for each predictive function. This measure then would be relevant to assessing either different predictive functions or various estimators of μ in terms of predictive discrepancy for the same predictive functions. We also note that other than the average D^{*}_{N-n} can be utilized e.g.: empirical distributions of the discrepancy can be compared for several predictors.

The answer to the second question is not at all obvious. Firstly, if we assume that the data setup adequately determines an observation we must first decide on n the number we shall omit. Once that is decided we then must choose a natural relevant partition. This is better described in an example. If we are dealing with J groups each having K independent observations and we may decide to omit only one observation at a time or we may decide to omit, say, as many as J observations at a time. If we decide on the first course, then the partition of the sample data would appear to be clear if every observation is to be treated symmetrically. But even in this case, i.e., one at a time, it may be desirable if the purpose is to predict only a new observation in a particular group to omit observations only in that group.

In the second instance omitting J at time may be done in a variety of ways; one way is to include every possible partition and another which may be more natural, is to consider only those partitions which omit simultaneously a single observation from each group, the latter, of course, considerably reduces the total number of partitions that are utilized.

If we are dealing with the mixed model J columns and K rows where we could reasonably consider the row as the observational unit then we might omit observations only in multiples of J. If we omit J observations, at a time, which is really only one vector observation with J components, the natural partition is reduced to K possibilities.

A clear and general choice of the schema is not really possible and a tentative selection can only be made for specific cases and even these cases may be hedged depending on the purposes of prediction and cross-validatory assessments.

The third issue involving predictive discrepancy measures assumes that the other two questions have been adequately decided if not fully resolved, but even then this is also subject to considerable equivocation. There are several possibilities. Firstly, D^{*}_{N-n} which is appropriate for comparative assessment, can be considered as a possibility when it utilizes the $\hat{\mu}_i$ using the same discrepancy function as for $\hat{\mu}$. This, when used a measure of the predictive discrepancy would probably tend to be conservative as it is based on N-n observations. However, some of the conservativeness may be offset due to repeated optimizations on much of the same data. On the other hand

$$D_N(\hat{\mu}) = P^{-1}n^{-1}\sum_{j\epsilon\Gamma} d(x_{jO}^{(n)}\ ,\ x(X,Y,y;\hat{\mu})), \tag{2.4}$$

based on N observations would inevitably yield more predictive precision than the data allow due to a single optimization on all of the data. A third possibility would be $D_{N,n}(\hat{\mu})$ which trades off a single optimization

with the fact that the predictor is based on N-n observations. All three measures then can have some relevancy in the assessment of the discrepancy of prediction. For some interesting problems explicit formulas can be obtained for $D_N(\hat{\mu})$ and $D_{N,n}(\hat{\mu})$ while D^*_{N-n} invariably requires a computer for its calculation. However, there are algebraic strictures on the first two measures that often make them less desirable than the latter.

As in the subjective Bayes approach there is much that is subjective here, but the prohibition that the subjectivity be prior or independent of the data is a dictum that falls by the wayside. The statistician interacts continually with the data in a variety of ways - trying out alternative schemata, predictive functions and discrepancy measures always assessing and testing until he feels he has exhausted the potentialities of the data and is satisfied with his predictor.

3. An Application to Several Groups.

Consider as an application the setup of J groups all measured on the same attribute with K_j observations in the j^{th} group $j = 1,\ldots,J$. The observations are x_{ij} where $i = 1,\ldots,K_j$. Here the index j takes the place of Y. In the usual analysis of variance parlance this is designated as either a fixed effect or a random effect model depending on certain sampling circumstances. It was already noted by Lindley [1971] that for the Bayesian such a distinction was essentially blurred although Box and Tiao [1968] in their Bayesian approach retained the distinction. We shall assume that the predictive function of interest is of the form $(1-\mu)\bar{x}_{.j} + \mu x_{..}$ for a future observation from the j^{th} group, where

$$\bar{x}_{.j} = K_j^{-1} \sum_k x_{kj} \; ; \; \bar{x}_{..} = N^{-1} \sum_{k,j} x_{kj} \; ; \; N = \sum_j K_j \; ; \; 0 \le \mu \le 1 \; .$$

This predictive function which is Stein type "shrinker" has some appeal when the population variation in each group is approximately the same and a certain affinity of the groups is assumed.

First we investigate a schema of one at a time omissions of the totality of observations and a discrepancy function of squared deviations. Hence, we write for the discrepancy function (we shall be using the same type of discrepancy function, but as the partition schema Γ will be varied the notation shall be changed for convenience)

$$s^2_{N,1}(\mu) = N^{-1} \sum_{j=1}^{J} \sum_{k=1}^{K_j} [(1 - \mu)c_{kj} + \mu\bar{c}_{kj} - x_{kj}]^2 \qquad (3.1)$$

where

$$c_{kj} = (K_j\bar{x}_{.j} - x_{kj})/(K_j - 1);\ \bar{c}_{kj} = (N\bar{x}_{..} - x_{kj})/(N-1).$$

Use of the usual identities of the analysis of variance leads to

$$s^2_{N,1}(\mu) = \frac{\mu^2 N}{(N-1)^2}\sum_{j=1}^{J} K_j(\bar{x}_{.j}-\bar{x}_{..})^2 + \sum_{j=1}^{J}\left[\frac{(1-\mu)^2K_j^2}{N(K_j-1)} + \frac{\mu^2N(K_j-1)}{(N-1)^2} + \frac{2\mu(1-\mu)K_j}{N-1}\right]s_j^2 \quad (3.2)$$

where

$$s_j^2 = (K_j - 1)^{-1}\sum_{d=1}^{K_j}(x_{kj} - \bar{x}_j)^2 .$$

In order to find the optimum μ we minimize $s^2_{N,1}(\mu)$ with respect to μ. This yields

$$\hat{\mu}_1 = \frac{(N-1)\sum_{j=1}^{J}\frac{K_j(N-K_j)s_j^2}{K_j-1}}{N^2\sum_{j=1}^{J}K_j(\bar{x}_{.j} - \bar{x}_{..})^2 + \sum_{j=1}^{J}\frac{(N-K_j)^2}{K_j-1}\ s_j^2} \quad (3.3)$$

so that the ensuing estimator is $\min[\hat{\mu}_1, 1]$.

For the special case $K_j = K$, presented previously by Geisser (1974), we obtain the following simplification

$$s^2_{N,1}(\mu) = \frac{K[JK-1-\mu(J-1)]^2}{(K-1)(JK-1)^2}\ m_1 + \frac{\mu^2JK(J-1)}{(JK-1)^2}\ m_2 \quad (3.4)$$

where

$$m_1 = J^{-1}(k-1)^{-1}\sum_{k=1}^{K}\sum_{j=1}^{J}(x_{kj} - \bar{x}_{.j})^2$$

$$m_2 = K(J-1)^{-1} \sum_{j=1}^{J} (\bar{x}_{.j} - \bar{x}_{..})^2 ,$$

and

$$\hat{\mu}_1 = \frac{(JK-1)m_1}{(J-1)m_1 + (K-1)Jm_2} . \tag{3.5}$$

A simple formula is also available here for the quantity $D_N(\hat{\mu})$ or in the present notation

$$s_N^2(\hat{\mu}) = N^{-1} \sum_j (K_j-1)s_j^2 + N^{-1} \hat{\mu}^2 \sum_j K_j(\bar{x}_{.j} - \bar{x}_{..})^2 . \tag{3.6}$$

Another schema for omissions in the case $K_j = K$ is to omit J observations simultaneously such that only one is left out of each group. This schema is symmetric in omissions only when $K_j = K$ for all j. The results here were presented previously, Geisser (1974), for a squared deviation discrepancy. In this case the estimator for μ is $\min(\hat{\mu}_2,1)$ where

$$\hat{\mu}_2 = \frac{Km_1}{(K-1)m_2 + m_1} \tag{3.7}$$

which results from the minimization of the discrepancy measure

$$t_{N,J}^2(\mu) = [\frac{(k-\mu)^2 + J^{-1}(2K - \mu)\mu}{K(K - 1)}] m_1 + (JK)^{-1}(J-1)\mu^2 m_2 . \tag{3.8}$$

A simple formula is also at hand for

$$t_N^2(\hat{\mu}) = \frac{K-1}{K} m_1 + \hat{\mu}^2(KJ)^{-1}(J-1) m_2 . \tag{3.9}$$

A third schema of omissions is to focus on say the j^{th} group (it may, for example, be that only prediction of a future observation from this group is desired) and omit observations from this group one at a time. The predictive function is, say, $(1-\alpha_j)\bar{x}._j + \alpha_j\bar{x}..$, for $0 \le \alpha_j \le 1$. The squared deviation discrepancy function is

$$u^2_{N,1}(\alpha_j) = K_j^{-1}\sum_{k=1}^{K_j}[(1-\alpha_j)c_{kj} + \alpha_j\bar{c}_{kj} - x_{kj}]^2 \qquad (3.10)$$

where c_{kj} and $\bar{c}_{kj}$ are defined as before. Here we obtain for the estimator, $\min[\hat{\alpha}_j,1]$, where

$$\hat{\alpha}_j = \frac{(N-1)(N-K_j)s_j^2}{(K_j-1)N^2(\bar{x}.. - \bar{x}._j)^2 + K_j^{-1}(N-K_j)^2s_j^2} \qquad (3.11)$$

is obtained from the minimization of

$$u^2_{N,1}(\alpha_j) = (K_j-1)^{-1}K_js_j^2\left[1 - \frac{\alpha_j(N-K_j)}{K_j(N-1)}\right]^2 + \frac{N^2}{(N-1)^2}(\bar{x}.. - \bar{x}._j)^2\alpha_j^2 \qquad (3.12)$$

with respect to α_j. We note that $\hat{\alpha}_j$ depends primarily on the variation within the j^{th} group and the squared deviation of the j^{th} group mean from the grand mean. When $K_j = K$, for all j

$$\hat{\alpha}_j = \frac{(JK-1)s_j^2}{(J-1)s_j^2 + (K-1)J^2(J-1)^{-1}K(\bar{X}.. - \bar{X}._j)^2} \qquad (3.13)$$

a form which is rather similar to $\hat{\mu}_1$ with s_j^2 replaced by m_1 and $J(J-1)^{-1}K(\bar{X}.. - \bar{X}._j)^2$ replaced by m_2.

A trivial computation reveals that $D_N(\hat{\mu}_{(j)})$ or in the present notation

$$u_N^2(\hat{\alpha}_j) = K_j^{-1}(K_j-1)s_j^2 + \hat{\alpha}_j^2(\bar{x}.. - \bar{x}._j)^2 \ . \tag{3.14}$$

It is also clear that simultaneous prediction for all $j = 1,\ldots,J$ merely involves minimizing the total discrepancy

$$u_{N,1}^2(\alpha) = u_{N,1}^2(\alpha_1,\ldots,\alpha_J) = N^{-1} \sum_{j=1}^{J} K_j u_{N,1}^2(\alpha_j) \tag{3.15}$$

the solution to which is the same as before, namely, $\hat{\alpha}_j$, $j = 1,\ldots,J$. Further

$$u_{N,1}^2(\hat{\alpha}_1,\ldots,\hat{\alpha}_j) \le s_{N,1}^2(\hat{\mu}_1) \tag{3.16}$$

since

$$u_{N,1}^2(\mu,\ldots,\mu) = s_{N,1}^2(\mu) \ .$$

The algebraic stricture (3.16) of course emphasizes the limited role that $D_{N,1}(\hat{\mu})$ has in general either as an actual measure of predictive discrepancy or in any comparision of alternative estimators of μ. Although it is often much easier to compute than D_{N-1}^* and may have considerable approximate value, great care must be exercised in its use.

4. Illustration

As an illustrative example of one of the previous applications of predictive sample reuse methodology we turn again to the Dyestuff Data previously discussed by Box and Tiao (1968), Novick (1971), and Geisser (1974).

Table I: Dyestuff Data

	Batch					
	1	2	3	4	5	6
	145	140	195	45	195	120
	40	155	150	40	230	55
	40	90	205	195	115	50
	120	160	110	65	235	80
	180	95	160	145	225	45
$\bar{x}_{\cdot j}$	105	128	164	98	200	70
s_j^2	3975	1107.5	1442.5	4720	2500	962.5

$J = 6$, $K = 5$, $m_1 = 2{,}451.25$, $m_2 = 11{,}271.50$, $\bar{x}_{\cdot\cdot} = 127.5$

TABLE II: Predicted Values (to the nearest integer)

	Method	1	2	3	4	5	6	$\hat{\mu}$	$s^2_{30,1}(\hat{\mu}_1)$	$t^2_{30,6}(\hat{\mu}_2)$	$u^2_{30,1}(\hat{\alpha})$
P	One-at-a-time$(\hat{\mu}_1)$	111	128	155	105	182	84	.251	2931	---	---
P_J	J-at-a-time $(\hat{\mu}_2)$	111	128	155	106	181	85	.258	---	2932	---
$P_($	Simultaneous single group $(\hat{\alpha}_j)$	128	128	156	126	193	73	---	---	---	2647
	$\hat{\alpha}_j$	1	1	.210	.919	.094	.058				

The closeness that methods P_1 and P_J yield in their predictions and the fact that for all intensive purpose $s^2_{30,1}(\hat{\mu}_1)$ and $t^2_{30,6}(\hat{\mu}_2)$ are the same indicate that there is little to choose in regard to these methods for this set of data. On the other hand P_G results in predictions that for 4 of the batches are substantially different from the others. The comparison of say, $s^2_{30,1}(\hat{\mu}_1)$ and $u^2_{30,1}(\hat{\alpha})$ indicates that the latter is appreciably smaller. Of course, we knew that it had to be smaller but the size of difference is indicative that a comparison of D^*_{N-1} for both these methods is at least in order before deciding which set of predictions one would prefer.

5. Flattening Simple Regression.

The preceding applications of the method have been in the realm of "skrinkers". Closely associated with shrinkers are "flatteners" and we now present some applications in this area to the same sorts of data bases as previously considered.

Suppose for sets of observations (x_{kj}, y_j) $k = 1,\ldots,K_j$; $j = 1,\ldots,J$ we specify a linear predictive function, which flattens the usual regression, as

$$x = \mu\bar{x}.. + (1-\mu)[\bar{x}.. + b(y-\bar{y})] \equiv \bar{x}.. + (1-\mu)b(y-\bar{y}) \tag{5.1}$$

where $\bar{x}.. = N^{-1} \sum_{j,k} x_{jk}$, $\bar{y} = N^{-1} \sum_{j=1}^{N} K_j y_j$ and

$$b = \frac{\sum_j K_j(\bar{x}._j - \bar{x}..)(y_j-\bar{y})}{\sum_j K_j(y_j-\bar{y})^2} = \frac{A}{B}, \tag{5.2}$$

with μ restricted to $[0, 1]$.

Assuming a squared deviation descrepancy with a one-at-a-time schema of omissions

$$s^2_{N,1}(\mu) = N^{-1} \sum_j [\bar{c}_{kj} + (1-\mu)\, b_{kj}(y_j-\bar{y}_j) - x_{kj}]^2 \tag{5.3}$$

where $\bar{c}_{kj}$ is defined as in section 3, $\bar{y}_j = (N-1)^{-1}(N\bar{y}-y_j)$ and b_{kj} is the usual regression coefficient omitting the observation x_{kj}. Simple algebra reveals that

$$b_{kj} = \frac{A - N(N-1)^{-1}(x_{kj}- \bar{x}..)(y_j-\bar{y})}{B - N(N-1)^{-1}(y_j-\bar{y})^2} \tag{5.4}$$

Minimization of (5.3) with respect to μ yields

$$1 - \hat{\mu} = \frac{\sum_j \sum_k (x_{kj} - \bar{x}..)(y_j - \bar{y})b_{kj}}{\sum_j \sum_k (y_j - \bar{y})^2 b_{kj}^2} = \hat{\beta}_1 \tag{5.5}$$

for the "flattener". The predictor then is

$$\hat{x} = \bar{x}.. + \hat{\beta}_1 b(y - \bar{y}) \ . \tag{5.6}$$

We now investigate another configuation wherein the observations $x_k' = (x_{k1}, \ldots, x_{kJ})$, $k = 1, \ldots, K$ for $K \geq 2$ are as in the mixed model, or perhaps it represents the k^{th} individual measured at J time points. We assume the same form for the predictive function as before namely

$$x = \bar{x}.. + (1-\mu)b(y - \bar{y}) \tag{5.6}$$

and a squared deviation discrepancy omitting the k^{th} vector i.e., J observations,

$$v_{N,J}^2(\mu) = N^{-1} \sum_{h=1}^{K} \sum_{j=1}^{J} (\bar{d}_k + (1-\mu)b_k(y_j - \bar{y}) - x_{kj})^2 \tag{5.7}$$

with $\bar{d}_k = (KJ-J)^{-1} [KJ\bar{x}.. - \sum_j x_{kj}]$ and

$$b_k = \frac{\sum_j (d_{kj} - \bar{d}_k)(y_j - \bar{y})}{\sum_j (y_j - \bar{y})^2} \tag{5.8}$$

where $d_{kj} = (K\bar{X}._j - x_{kj})/(K-1)$. Equation (5.8) is equivalent to

$$b_k = (K-1)^{-1}(Kb - b_k') \tag{5.9}$$

where

$$b_k' = \frac{\sum_j (x_{kj} - \bar{x}_{k.})(y_j - \bar{y})}{\sum_j (y_j - \bar{y})^2} \tag{5.10}$$

i.e., the usual regression coefficient of the k^{th} row upon $y_1,\ldots,y_J$.

Minimization of (5.7) with respect to μ yields, after some trivial algebra, the flattener

$$1-\hat{\mu} = \frac{\sum_k b\, b_k'}{\sum_k b_k^2} = \frac{(K-1)[K^2b^2 - \sum_k b_k'^2]}{K^2(K-2)\, b^2 + \sum_k b_k'^2} = \hat{\beta}^* \tag{5.11}$$

so that the predictor is

$$\hat{x} = \bar{x}.. + \hat{\beta}^* b(y-\bar{y}) \ . \tag{5.12}$$

It is appropriate to point out that for both of these cases when $K_j = K$ we could approach flattening via the route of "shrinkers". In section 3 we treated the first data base in the context of shrinking towards the grand mean. If we then computed the least squares regression thru the shrunken "means" we would easily obtain

$$\hat{x} = \bar{x}.. + (1 - \hat{\mu})b(y - y_j) \tag{5.13}$$

Here for $\hat{\mu}$ we would substitute for one-at-a-time omissions

$$\hat{\mu}_1 = \frac{(JK-1)m_1}{(J-1)m_1 + (K-1)Jm_2} \tag{5.14}$$

and for J at-a-time omissions (the second schema of section 3)

$$\hat{\mu}_2 = \frac{Km_1}{m_1 + (K-1)m_2} . \tag{5.15}$$

Shrinking mixed model means is presented in Geisser (1974) and for that case we would substitute for μ,

$$\mu^* = \frac{K(m_1 J - m_3)}{(J-1)(K-1)m_2 + Jm_1 - m_3} = \frac{Km_I}{(K-1)m_2 + m_I} \tag{5.16}$$

where

$$m_3 = J(K-1)^{-1} \sum_{u=1}^{K} (\bar{x}_{k.} - \bar{x}..)^2 , \quad \bar{x}_{k.} = J^{-1} \sum_{j=1}^{J} x_{kj} ,$$

and m_I represents the mean square for interaction.

Another possibly "natural" schema is to omit columns one-at-a-time. This omits K observations at-a-time in a selective way. For squared discrepancy this yields the predictor

$$x = \bar{x}.. + \hat{\beta}^{**} b(y - \bar{y}) \tag{5.17}$$

where

$$\hat{\beta}^{**} = \frac{\sum_j (\bar{x}._j - \bar{x}..)(y_j - \bar{y}) b_{(j)}}{\sum_j (y_j - \bar{y})^2 b_{(j)}^2} \tag{5.18}$$

and $b_{(j)}$ is the usual regression of $\bar{x}._i$ on y_i, $i = 1,\ldots,J$ but $i \neq j$.

We conclude this section by deploying the method in a situation where we base a prediction on a simple combination of regressions. We deal with the mixed model paradigm i.e., K individuals measured at the same J time points. In addition we have the first $J-1$ observations on a $K+1^{st}$ individual and our goal is to predict his value at the, as yet, unobserved J^{th} time point. We wish to combine the regressions from the original K individuals and the $K+1^{st}$ to predict the value in question. Let the predictor be

$$x_{K+1,J} = (1-\mu)\,\tilde{x}_{K+1,J} + \mu\,\tilde{x}_J \tag{5.14}$$

where

$$\tilde{x}_{K+1,J} = d_{K+1,J} + b_{(K+1,J)}(y_J - \bar{y}_J) \tag{5.15}$$

and d_{kj} is as previously, the mean of the elements of the k^{th} vector with the j^{th} component omitted and $b_{(kJ)}$ is the usual repression coefficient based on the k^{th} vector with (x_{kJ}, y_J) omitted;

$$\tilde{x}_J = \bar{x}.. + b(y_J - \bar{y}), \tag{5.16}$$

essentially is a predictor derived from the usual regression based on the first K individuals.

Assuming quadratic descrepancy and one-at-a-time omissions from the J^{th} column

$$D_{N,1}(\mu) = K^{-1} \sum_{k=1}^{K} [(1-\mu)\,\tilde{x}_{kJ} + \mu\,\tilde{x}_{(k)J} - x_{kJ}]^2 \tag{5.12}$$

where $\tilde{x}_{(k)J}$ is as in (5.16) but with the k^{th} vector omitted. Minimization of $D_{N,1}(\mu)$ with respect to μ yields

$$\hat{\mu} = \frac{\sum_{k=1}^{K}(\tilde{x}_{kJ} - x_{kJ})(\tilde{x}_{kJ} - \tilde{x}_{(k)J})}{\sum_{k=1}^{K}(\tilde{x}_{kJ} - \tilde{x}_{(k)J})^2} \quad .$$

Of course J plays no vital role in the algebra and it could just as well be replaced by any single $j = 1, 2, \ldots J-1$. In such a case one would be "predicting" an unrecorded past value in a time series. If something other than a time series were involved, then obviously J should be replaced with j.

This can easily be extended, under suitable restrictions on K and J, in three directions, firstly for other functional relationships and secondly for more than one predicted value per new individual, and thirdly for several new individuals. This more general problem referred to as "partial prediction" is discussed in the highly structured case by Lee and Geisser (1972).

6. Predicting future successes.

Suppose we have a sequence of Bernoulli trials say $x_1,\ldots,x_N$ $x_i = 0$ or 1, and we wish to predict the number of successes in the next M Bernoulli trials. Assume that the observed first N trials resulted in r successes. We let the predictive function for the number of successes be

$$f = \frac{M(r + \mu)}{N + 2\mu} \tag{6.1}$$

for $\mu \geq 0$ suggested by Bayesian analysis. Using a squared discrepancy function with a one-at-a-time omission schema

$$s^2_{N,1} = N^{-1} \sum_{j=1}^{n} \left[\frac{M(r - x_j + \mu)}{N - 1 + 2\mu} - Mx_j\right]^2 . \tag{6.2}$$

Evaluation of (6.2) yields

$$s^2_{N,1} = \frac{N^{-1}M^2[N\mu^2 + 4r(N-r)\mu + rN(N-r)]}{(N - 1 + 2\mu)^2} . \tag{6.3}$$

Minimization of the above with respect to μ, for $\mu \geq 0$, yields for μ

$$\left.\begin{array}{ll} \dfrac{2r(N-r)}{N(N-1) - 4r(N-r)} & \text{if } N(N-1) > 4r(N-r) \\ \infty & \text{otherwise .} \end{array}\right\} \tag{6.4}$$

Hence

$$\left.\begin{aligned} \hat{f} &= \frac{Mr[N(N-1) - 2(N-r)(2r-1)]}{(N-1)(N-2r)^2} \qquad \text{if } N(N-1) > 4r(N-r) \\ \hat{f} &= \frac{M}{2} \qquad \text{otherwise .} \end{aligned}\right\} \tag{6.5}$$

This predictor has the property of almost always being closer to $\frac{M}{2}$ than the "natural"predictor $\frac{Mr}{N}$, i.e.,

$$|\hat{f} - \tfrac{M}{2}| \leq |\tfrac{Mr}{N} - \tfrac{M}{2}| \ . \tag{6.6}$$

It is of interest to point out that this predictor can also be derived by a method of "marginal moments" Sutherland et al,(1974) which is based on a suggestion made by Good (1965).

If for this problem, we minimize the absolute deviation discrepancy

$$\begin{aligned} S_{N,1} &= N^{-1} \sum_{j=1}^{N} \left|\frac{M(r - x_j + \mu)}{N - 1 + 2\mu} - M x_j\right| \\ &= N^{-1} M \frac{(N\mu + 2r(N-r))}{N - 1 + 2\mu}, \end{aligned} \tag{6.7}$$

we obtain solutions for $\mu \geq 0$

$$\begin{aligned} \mu &= 0 \qquad \text{if } 4r(N-r) \leq N(N-1) \\ \mu &= \infty \qquad \text{if } 4r(N-r) > N(N-1) \ . \end{aligned}$$

We note that when $4r(N-r) = N(N-1)$, $S_{N,1}$ is independent of μ and

hence it is immaterial which value of μ is used. Therefore

$$\left.\begin{aligned} \hat{f} &= \frac{M\,r}{N} \qquad \text{if} \quad 4r(N-r) \le N(N-1) \\ \hat{f} &= \frac{M}{2} \qquad \text{if} \quad 4r(N-r) > N(N-1) \ . \end{aligned}\right\} \tag{6.8}$$

This predictor then is the usual predictor except when r is close to $\frac{M}{2}$ and there it becomes $\frac{M}{2}$ and hence retains the "centering" property (6.6).

We now present results for t-at-a-time omissions. For the absolute discrepancy we obtain

$$\begin{aligned} \hat{f}_{1t} &= \frac{M\,r}{N} \qquad \text{if} \quad 4r(N-r)(N-t+1) \le N^2(N-t) \\ \hat{f}_{1t} &= \frac{M}{2} \qquad \text{otherwise} \ . \end{aligned} \tag{6.9}$$

where the subscript 1 refers to absolute discrepancy and t the number of omissions. We also note that the centering property is stronger as the number of omissions is increased, i.e.,

$$|\hat{f}_{1,t+1} - \tfrac{M}{2}| \le |\hat{f}_{1t} - \tfrac{M}{2}| \tag{6.10}$$

for $t = 1, \ldots, N-2$.

For the squared discrepancy we obtain similarly

$$\left.\begin{aligned} \hat{f}_{2t} &= \frac{Mr[N^2(N-t)-2(N-r)(2r(N-t+1)-N)]}{N(N-t)(N-2r)^2} \qquad \text{if} \quad 4r(N-r)(N-t+1) \le N^2(N-t) \\ \hat{f}_{2t} &= \frac{M}{2} \qquad \text{otherwise} \ . \end{aligned}\right\} \tag{6.11}$$

Again it can easily be shown that

$$|\hat{f}_{2,t+1} - \frac{M}{2}| \leq |\hat{f}_{2t} - \frac{M}{2}| \ . \tag{6.12}$$

In addition it is also clear that

$$|\hat{f}_{2t} - \frac{M}{2}| \leq |\hat{f}_{1t} - \frac{M}{2}| \tag{6.13}$$

for each $t = 1,\ldots, N-1$. Hence use of the squared discrepancy yields a stronger "centerer" than the absolute discrepancy for each t.

In the extreme case, where $t = N-1$, $N \leq 6$,

$$\left.\begin{aligned} \hat{f}_{1t} = \hat{f}_{2t} &= 0 && \text{for } r = 0 \\ &= \frac{M}{2} && \text{for } 1 \leq r \leq N-1 \\ &= M && \text{for } r = N \end{aligned}\right\} \tag{6.14}$$

maximal centering is achieved.

The predictive function (6.1) was utilized to display the "centering" phenomenon. This is actually a flattening of values of the predictor around the center $M/2$. If for some a priori reason one wished to effectuate a flattening around some other value, say, M/a, for some known $a \geq 1$, one need only alter the predictive function to

$$f(a) = \frac{M(r + \mu)}{N + a\mu} \ . \tag{6.15}$$

One can then easily obtain $\hat{f}_{1t}(a)$ and $\hat{f}_{2t}(a)$ in the same manner as previously, suitably flattened about M/a.

7. Summary.

A low structure predictivistic approach has been delineated, that simulates as best it can the most fundamental process of inference, prediction. Although highly flexible and versatile, it engenders problems of its own, primarily in selection and assessment. It is intended to serve as a complement to the tightly structured Bayes apparatus of prior cum likelihood that will yield the whole spectrum of possible values and associated probabilities, subsumed in the rubric, predictive distribution. The Predictive Sample Reuse method assuming less yields less.

The predictivistic view is not so much a mode of inference as it is an operational dichotomy, focussing on objectives rather than logical distinctions. In its extreme form it implies that inferences be restricted to entities that are at least potentially observable unless there are strong reasons to the contrary.

References

Box, G. E. P. and Tiao, G. (1968). Bayesian estimation of means for the random effect model. JASA, Vol. 63, pp. 174-181.

Geisser, S. (1971). The inferential use of predictive distributions, Foundations of Statistical Inference, Godambe and Sprott, eds. Holt, Rinehart and Winston, pp. 456-469.

Geisser, S. (1974). A predictive approach to the random effect model. Biometrika, 61 (to appear).

Good, I. J. (1965). The Estimation of Probabilities, Cambridge, Mass., Massachusetts Institute of Technology Press.

Lee, J. C. and Geisser, S. (1972). Growth Curve Prediction, Sankhyā, Series A, No. 4, p. 393-412.

Lindley, D. V. (1971). Bayesian statistics, a review, SIAM, Regional Conference Series in Applied Mathematics, No. 2.

Mosteller, F. and Tukey, S. W. (1968). Data analysis including statistics. In Handbook of Social Psychology (G. Lindzey and E. Aronson, eds.), Vol. 2, Reading, Mass: addison-Wesley.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. JRSS, Series B, Vol. 36 (to appear).

Sutherland, M., Holland, P. W., and Feinberg, S. E., (1974). Combining Bayes and Frequency Approaches to Estimate a Multinomial Parameter, in S. E. Fienberg and A. Zellner, eds., Studies in Bayesian Econometrics and Statistics, Amsterdam: North Holland Publishing Co. (to appear).